Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 12 (2012) 93 – 98

Procedia Environmental Sciences

2011 International Conference on Environmental Science and Engineering (ICESE2011)

# Douhe Reservoir Flood Forecasting Model Based on Data Mining Technology

He Ji [1], Wang Songlin [1], Wu Qinglin [2], Chen Xiaonan [3]

[1]North China University of Water Resources and Electric Power, Zhengzhou 450011
[2]Douhe Reservoir Management Department, Tangshan 063000
[3]South-to-North Water Diversion Program (mid-line) Construction Management Bureau, Beijing 100038

**Abstract**

Calculating flood based on rainfall is an important part of hydrological forecast. However, due to the diversity and complexity of factors affecting the relationship between rainfall and runoffs, using the perspective of mechanism to simulate the forming of flood through rainfall is often difficult. In this paper, flood forecast model is constructed based on Artificial Neural Networks (ANN) and Genetic Programming (GP), using actual data to mine the relationship among rainfall, pre rain and net rain, to avoid the flaws of constructing actual mathematical expression in advance, and automatically search for optimal structure. Practice has approved that applying data mining technique on flood forecasting of Douhe Reservoir is able to achieve outstanding results.



*Keyword:*Hydrological Forecasting, Data Mining Technology, Artificial Neural Networks

Calculating flood based on rainfall is an important part of hydrological forecast. However, due to the diversity and complexity of factors affecting the relationship between rainfall and runoffs, using the perspective of mechanism to simulate the forming of flood through rainfall is often difficult because of the requirement of large actual data. So far a Storm-Runoff relation chart is commonly used. This approach uses pre-rainfall parametres and actual data to relate storms and run off through charting, hence suffers human perspective and subjectivity.

Data Mining is an advanced information processing technology which discovers laws over data to obtain useful information. In broad sense, any method that extracts information from data can be regarded as data mining, which includes a variety of information processing methods. This study uses Artificial Neural networks and Genetic Programming approaches to mine the relationship between rainfall and runoff.

 doi:10.1016/j.proenv.2012.01.252

## 1. Artificial Neural Networks

### *1.1. BP (Back Propagation)*

Artificial Neural Network is based on the human brain and its activities, to establish a theoretical mathematical model in terms of bionics. The model uses physical simulation of human thoughts and intellectuals. Among researches applied in hydrological researches, the mostly wide used is BP network.

Fu Qiang from North-Eastern Agricultural University uses self-organizing competitive neural network to classify the soil [1]. Zhang Libing from Hefei Industrial University chooses better prediction method using BP network. Zhang Xiang from Wuhan University use modified BP network to forecast Flood. Ding Jing from Sichuan University uses sensitive neural network in hydraulic forecasts.

### *1.2. BP network based on Simulated Annealing*

The fundamental idea of Simulated Annealing originates from the Annealing Process of solid matter in Physics. It is often used to solve combination optimization problems. In this paper Simulated Annealing is combined with BP network to overcome its weakness of the slow convergence, in the meantime ensures the quick finding of global minimum.

Neural network structure consists of three layers including input layer, hidden layer and output layer. Applications generally assume that the input data an n dimensional vector, the output data an m dimensional vector, and the hidden layer contains h neurons.

**1.2.1. BP network leaning process**

Using W and V accordingly as the weight matrix of output layer, input (hidden) layer, $w_{ij}$ ( $i = 1,2,\cdots,h, j = 1,2,\cdots,m$ ) represents the weight value between the ith neuron in the hidden layer and the jth neuron in the output layer, $v_{ij}$ ( $i = 1,2,\cdots,n, j = 1,2,\cdots,h$ ) represents the weight value between the ith neuron in the input layer and the jth neuron in the hidden layer.

$(X,Y)$ represents a sample in a sample concentration, where in $X = (x_1, x_2, \cdots, x_n)$ , $Y = (y_1, y_2, \cdots, y_m)$ . And the activation function is：

$$f(net) = \frac{1}{1+e^{-net}} \tag{1}$$

The actual output in corresponding to the sample is： $O = (o_1, o_2, \cdots, o_m)$ ， the output of the hidden layer is： $O = (o_1^{'}, o_2^{'}, \cdots, o_m^{'})$ 。

Where in

$$o_j = f(net_j) = f\left(\sum_{i=1}^{h} w_{ij} \cdot o_i^{'}\right) \quad (2) \quad j = 1,2,\cdots,m$$

$$o_j^{'} = f(net_j^{'}) = f\left(\sum_{i=1}^{n} v_{ij} \cdot x_i\right) \quad (3) \quad j = 1,2,\cdots,h$$

The error metric of any sample is：

$$E = \frac{1}{2}\sum_{k=1}^{m}(y_k - o_k)^2 \tag{4}$$

$\sum E$ is the error metric of the sample concentration.

Then the delta $\Delta w_{ij}$ of the weight value in the output layer is:

$$\Delta w_{ij} = \alpha \delta_j o_i^{'} = \alpha \cdot \left(y_j - o_j\right) \cdot \left(1 - o_j\right) \cdot o_j \cdot o_i^{'} \quad (5)$$

Where in $\alpha$ is the learning efficiency.

The delta of the connection weights in the hidden layer is:

$$\Delta v_{ij} = \alpha \cdot \sum_{k=1}^{m} (\delta_k \cdot w_{jk}) \cdot \left(1 - o_j^{'}\right) \cdot o_j^{'} \cdot x_i$$

Make $\sum_{k=1}^{m} (\delta_k \cdot w_{jk}) \cdot \left(1 - o_j^{'}\right) \cdot o_j^{'} = \delta_j^{'}$

Then $\Delta v_{ij} = \alpha \cdot \delta_j^{'} \cdot x_i$ (6)

**1.2.2. Adjusting the network weight value**

This study uses simulated annealing algorithm to adjust weights, and four steps are included.

(1) Initialize the V and W in every layer and define $T_0$ as the default value of artificial temperature.

(2) For every T, repeat the following work:

Choose tolerance value $E$ as the sample calculation deviation; Use Cauchy allocation to value every $w_{ij}$ and $v_{ij}$ with $\Delta w_{ij}$、$\Delta v_{ij}$; The recalculate tolerance value E with new weight value, then minus of the old and new E to get $\Delta E$; if $\Delta E \le 0$, $\Delta w_{ij}$、$\Delta v_{ij}$ is accepted, if $\Delta E > 0$, the choose random r within [0,1] using uniform distribution, also use $P = \exp\left(-\frac{\Delta E}{kT}\right)$ to calculate acceptance percentage ($k$ is Boltzmann constant), if $p < r$, $\Delta w_{ij}$、$\Delta v_{ij}$ is not accepted, continue calculation until the new $\Delta w_{ij}$、$\Delta v_{ij}$ value is accepted according to probability.

(3) Calculate reducing $T$.

(4) If $T$ is relatively large, calculate according to step 2, until T is small enough to finish calculation.

## 2. Genetic Programming Design

Genetic Programming (GP) is a kind of automation programming technique which is a new derivation of algorithm based on genetic algorithm. Being superior in highly accurate and adaptive, it excels at automatic model structure search, and is applied more and more in the fields of model constructing, forecasting and the design of neural networks [49].

The following steps are necessary in the appliance of genetic programming design:

Determine the objective function.

Assume $X = \left\{\left(x_{11}, x_{12}, \cdots, x_{1m}\right), \left(x_{21}, x_{22}, \cdots, x_{2m}\right), \cdots, \left(x_{n1}, x_{n2}, \cdots, x_{nm}\right)\right\}$, $Y = \left\{y_1, y_2 \cdots, y_n\right\}$ are input and output sample series. The design process is to determine the best expression of function $G\left(c, x_1, \cdots, x_m\right)$, to minimize the error.

$$\min f = \sum_{k=1}^{n} \left| G(c, x_{k1}, \cdots, x_{km}) - y_k \right| \quad (7)$$

c in function (7) is a real number constant.

（2）Coding. Determine the terminal set T (in which elements are defined as variable $x$ and constant c) and function set F (in which elements are basic {+，-，×，/}) and basic functions {sin，cos，arctg，arcctg}, and encode the union D = T ∪ F. Table 2.1 provides a recommended coding scheme. Function $G(c, x_1, \cdots, x_m)$ in genetic programming is expressed as a binary tree form and the solution of equation (7) .

Table.1 One coding scheme of GP

| Elements in D | c | x | + | - | × | / | sin | cos | arctg | arcctg |
|---|---|---|---|---|---|---|---|---|---|---|
| Code Value | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

（3）Initialize parent population. Set N as population size, and use the small integer as the maximum depth of trees, such as 4 or 6, to analyze and interpret the optimized function expression found. The root node of every single parent population can be randomly chosen from the code values corresponding to F, meanwhile intermediate node can be chosen from the code values corresponding to D, and finally leaf nodes can be chosen from the code values from T.

（4）Decode parent population and conduct adaption evaluation. Put G into (7) to get $f(i)$（i=1, 2,…N），and order these subjective function values from small to large, and individuals in front are outstanding individual. It can be defined the function value of the ith parent individual after adaption ordering as follows:

$$F(i) = 1/\left[ f(i) \times f(i) + 0.001 \right] \quad (8)$$

（5）Apply genetic manipulation of parent population, then the chance of parent individual i being selected is:

$$s(i) = F(i) / \sum_{i=1}^{N} F(i) \quad (9)$$

Make $p(0) = 0, p(i) = \sum_{k=1}^{i} s(k), i = 1,2,\cdots,N$。For the even random number $\mu$ generated, if it is in [ $p(i-1)$ , $p(i)$ ], then select the ith individual.

Crossover operation and mutation operation are often carried out in genetic programming. Crossover operation is to randomly create two crossing nodes and exchange them with offspring tree while mutation operation selects a randomly generated mutation node and use it as a root node, and exchange it with the offspring tree under it in accordance to step 3.

（6）Record the best offspring individuals as the new parent group and go back to step 4, repeat evolution until the number exceeds default value or subjects function value reaches default, and the outstanding individuals are the final result.

## 3. Douhe Reservoir Flood Forecasting Model Based on Data Mining Technology

### *3.1.Summary of Douhe Reservoir*

Douhe Reservoir is located upstream the River Dou, 15km north east to Tangshan City. It is a massive composite hydraulic project mainly used in flood control as well as providing citizens in Tangshan with living water and agricultural water. The River Dou is a seasonal flood river, and flood control area of the reservoir contains not only the whole Tangshan City, but also 1 million mu of farm land, transportations routes such as Jingshan Rail Line, Jingha Express Way, Jingtang Express Way and State Road 205, factories such as Tangshan Iron and Steel Company, Qinxin Cement Plant, Jianzhu Ceramics Company, Huaxin Textile Mill and Douhe Power Plant. Flood control is crucial.

Table 2 Actual Data of rainfall- runoff in Douhe Reservoir

| Peak No. | Average Rainfall(mm) | Average Pre-rain(mm) | Actual Runoff(mm) |
|---|---|---|---|
| 56904 | 61.4 | 38.7 | 4.8 |
| 59721 | 250.7 | 55.6 | 95.1 |
| 59801 | 69.3 | 77.6 | 24.0 |
| 59818 | 85.9 | 39.3 | 20.7 |
| 61719 | 79.0 | 75.4 | 18.9 |
| 64618 | 74.3 | 45.6 | 10.5 |
| 64706 | 75.1 | 27.9 | 4.9 |
| 65707 | 30.0 | 22.4 | 4.1 |
| 65710 | 34.2 | 37.0 | 1.8 |
| 66728 | 59.1 | 17.1 | 2.3 |
| 66814 | 55.6 | 39.4 | 2.1 |
| 66830 | 49.9 | 75.5 | 8.6 |
| 67624 | 62.5 | 30.1 | 5.0 |
| 67820 | 87.4 | 44.0 | 9.8 |
| 68818 | 86.8 | 22.2 | 3.6 |
| 69811 | 122.7 | 48.4 | 20.7 |
| 69816 | 96.3 | 59.0 | 24.9 |
| 70721 | 74.8 | 58.3 | 10.0 |
| 70729 | 8.1 | 61.2 | 1.2 |
| 70808 | 79.9 | 50.1 | 10.9 |
| 71721 | 14.2 | 53.0 | 0.9 |
| 72804 | 135.4 | 32.7 | 24.3 |
| 73820 | 78.3 | 43.6 | 4.3 |
| 74808 | 93.5 | 54.0 | 9.0 |
| 75730 | 192.1 | 16.1 | 11.6 |
| 75811 | 183.9 | 33.3 | 27.3 |
| 76723 | 149.5 | 37.8 | 15.5 |
| 77723 | 120.9 | 50.5 | 17.3 |
| 77726 | 186.9 | 73.4 | 51.4 |
| 78728 | 62.1 | 77.5 | 10.9 |
| 78806 | 121.4 | 44.5 | 10.8 |
| 79728 | 79.1 | 65.3 | 9.9 |
| 79814 | 111.7 | 66.2 | 26.1 |

### *3.2. Flood Forecast of Douhe Reservoir*

Runoff is derived from the average rain fall in the area. Using average rain fall value and the pre rain value in the area as the input of the model, and run off as the output of the model, the model in accordance with ANN is double input single output. The hidden layer neutron number is experimented and set to 10. The related historical rain fall-runoff data of Douhe Reservoir is as follows:

Assume the maximum value in every column Max and minimal Min, the following formula is used to normalize data:

$$v = (a - \min)/(\max - \min) \quad (10)$$

Where in： $a$ is original data， $v$ is normalized data

After calculation according to Artificial Neural Network , the input layer and the output layer matrix of the neural network are:

$X$ =[1.26，-0.56，1.35，0.27，-2.03，-0.82，-1.19，-0.74，-0.65，0.79
0.32，-2.04，-0.06，-0.40，1.62，-1.75，-0.75，-1.51，-1.64，1.40]

$Y$ =[2.00，-2.22，1.71，0.05，-4.12，-3.67，-4.39，-2.07，-1.90，1.83]T

Genetic Programming gets the following equation:

$F(x_1, x_2) = arctg\ (arcctg\ (x_2 + 0.061)) \times (x_1\ /(\cos\ x_1 + 0.0001)) \times x_2$ Where in $x_1$ and $x_2$ is the normalized result of average rainfall and pre rain. The two models have a fitting accuracy of 0.013 and 0.0128.

## 4. Appliance Analysis

With the rapid development of information technology, data accumulation is also increasing sharply, in the size of tera Bytes. The way to extract useful information from the database is the paramount requirement. Data Mining is the data processing techniques developed in comply with this need. In this paper, artificial neural networks and genetic programming is used to establish flood forecasting models. It is based on actual data to mine the relationship among rainfall pre rain and net rain, in order to avoid the lacking in previously established mathematical expressions. It is able to find the optimal structure automatically and yields good accuracy. Utilizing data mining technology to establish the flood forecast model for Douhe reservoir is proved to achieve excellent results.

## References

[1] Fu Qiang,Wang zhiliang. Application of Self—Organizing Competition Artificial Neural Networks in Soil Classificati0n. Bulletin of Soil and Water Conservation. 2002, (1):39~43.

[2] Zhang libing,Jinjuliang. The forecasting method of selecting-best model based on artificial neural networks and its application. Journal of Hydroelectric Engineering.. 2007, (6):12~16.

[3] Zhang xiang,Song xingyuan. Variable Structure Neural Network for Short Term Flood Forecasting. International Journal Hydroelectric Energy. 2002, (1):12~14.

[4] Ding Jing,Qin Guanghua.. Application of an ANNs with sensitive ability to hydrologic forecast. Advances in Water Science.2003, (2):163~166.

[5] Chen Xiaonan,Qiu Lin. The BP Network Based on Simulated Annealing Algorithm and Application in Hydrology. Journal of North China Institute of Water Conservancy and Hydroelectric Power[J] .2005, (1):1~3.